ABSTRACT

Data mining is a valuable business tool that companies can utilize to understand their customers and attain competitive advantage. A critical component of data mining is classifier selection; companies must meticulously select an appropriate classifier as it impacts the accuracy of the results. In order to select an appropriate classifier, a company's knowledge discovery team must master a lot of background information of the dataset, the model and the algorithms in question. We suggest that recommender systems can ease this complex process by searching the knowledge stored in the result repository and recommending an appropriate classifier to be used for a particular dataset. In this study we propose such a system and take a first look on how it can be done. We compare various classifiers against different datasets and then come up with the most appropriate classifier for a particular dataset based on its unique characteristic. The results of our experiments indicate that AdaBoost is a relatively stable performer compared to other algorithms. Other findings and managerial implications are also discussed in our study.

Key Words: *Data Mining, Algorithm Selection, Model Selection, UCI, Weka, Recommender Systems*

Data mining in business Domains:

A conceptual model of Recommender Systems for classifier Selection

Mujtaba Ahsan

INTRODUCTION

Data mining is a valuable technology that managers can use when making complex decisions. Fayyad et al. (1998) describe data mining as one of the primary phases in the knowledge discovery process, as it extracts useful patterns from databases. Data mining has increasingly been utilized in diverse domains such as bankruptcy prediction (Donato et al., 1999), traffic safety programs (Solomon et al., 2006), customized and personalized marketing (Shen and Chuang, 2009; Pitta, 1998) and customer service support (Hui and Jha, 2000). Similarly, Adriaans and Zantinge (1996) and Bigus (1996) provide a fundamental concept for the utilization of data mining in business problems covering customer ranking, marketing segmentation, real estate pricing, sales forecasting, and customer profiling. With the growing popularity of data mining as indicated by the above-mentioned literature, it is important that critical aspects of the data mining process are highlighted so that organizations become aware of them. One such critical aspect of the data mining process is model and algorithm selection. In order to undertake a data mining process the knowledge discovery team has to first select an appropriate model and algorithm (i.e., classifier) for making a prediction or classification. Simply put, a classifier is a data mining statistical technique used for analyzing the dataset and making predictions. This selection is probably one of the most difficult problems in data mining, since there is no model or algorithm that is better than all others independent of the particular problem characteristic (Aha, 1992; Salzberg, 1991; Shavlik, Mooney and Towell, 1991; Weis and Kapouleas, 1989). Each algorithm has a certain distinct advantage (Brodley, 1995) (i.e. the algorithm in question, under certain conditions or for specific types of problems is better than the rest). This happens because every algorithm has an "inductive bias" (Mitchel, 1997) caused by the assumptions it makes in order to generalize from the training data to the unknown data Hence, the knowledge discovery team must posses a lot of experience to be able to identify the most appropriate algorithm for the morphology of the problem at hand.

Another important function of data mining is the production of a model. A model can either be descriptive or predictive. A descriptive model helps in understanding the underlying processes or behavior, whereas, a predictive model is an equation or set of rules that makes it possible to predict an unseen or unmeasured value (the dependent variable) from other, known values (independent variables). The form of the equation or rules is suggested by mining data collected from the process under study. Some training or estimation technique is used to estimate the parameters of the equation or rules. It should be noted that the model of an algorithm actually defines the area of search, such as rules, k-DNF forms, linear discriminant functions, etc. Using this information the algorithm searches the defined space for a hypothesis that best fits the data by determining the order of search. For example, one algorithm might start the search in an area that contains the complete set of features, whereas the other algorithm might start search in an area that consists of only one feature (Gordon and desJardin, 1995). The wrong choice of algorithm may result in a slow convergence or suboptimal solution. Also, choosing the wrong model can result in an appropriate hypothesis for the problem at hand being ignored as it is not contained in the model's search space (Gordon and desJardin, 1995). This suggests that selecting an appropriate model and algorithm is critical in enhancing the knowledge discovery process. Thus, in this paper we look at this selection process based on the morphology and special characteristics of the problem at hand. The paper further addresses the development of a prototype repository and tools that can be utilized to select appropriate models and algorithms.

number of results already exist in the repository (F). Also its applicability might be more relevant to organizations that utilize the data mining process frequently. We conducted some experiments to test for the accuracy of various classifiers on different sets and have described the experiments in detail in the next section.

Experiments

For the purpose of an initial study, we selected six different UCI datasets whose characteristics are shown in Table 1 and focus our study on algorithm selection. In order to compare the performances of the various algorithms, we performed experiments on a collection of six data sets from the UCI Repository. We selected the six data sets based on the following criteria

- Number of classes (i.e. two-class or multiple-class)
- Types of features (i.e. Nominal, Numeric, or both)

Table 1: Characteristics of 6 UCI Data Sets

No.	Name	Classes	Instances	Features		
				Nominal	Numeric	Total
1	Audiology	24	226	69	0	69
2	Automobile	7	205	10	16	26
3	Glass Identification	7	214	0	9	9
4	Horse Colic	2	368	15	7	22
5	Ionosphere	2	351	0	34	34
6	Breast Cancer	2	286	9	0	9

The experiments were run utilizing the Weka data mining tool (Witten and Frank, 2000). The next step was the algorithm selection, where the selection process was limited to those algorithms available in the Weka suite. The following algorithms were utilized in the experiments: the decision tree learning algorithm J48, which is a re-implementation of C4.5, the k-nearest neighbor (IBk) algorithm, the naive Bayes (NB) algorithm, the neural network algorithm and the ZeroR algorithm. Also three algorithms for combining classifiers, namely bagging, boosting and stacking were included. The performance of each of these algorithms is assessed in terms of its accuracy and error rate. In all experiments, classification errors are estimated using 10-fold stratified cross validation. Cross validation is repeated ten times using different random generator seeds resulting in ten different sets of folds (Kohavi, 1995).

RESULTS

The results of the various experiments were analyzed on the following two aspects

- **Accuracy** - Accuracy is an important factor in assessing the success of data mining. When applied to data, accuracy refers to the rate of correct values in the data. When applied to models, accuracy refers to the degree of fit between the model and the data. This measures how error-free the model's predictions are. These results can be seen in Table 2.
- **Confusion Matrix** - A confusion matrix shows the counts of the actual versus predicted class values. It shows not only how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong. Based on the confusion matrix, we can calculate true positive, false positive, true negative and false negative (Sinha and May, 2005). These results can be seen in Table 3.

Table 2: Classifiers Accuracy

Datasets	NN	Bagging	AdaBoost	Stacking	J48	Ibk	Naive Bayes	Zero R
Audiology	83.2964	81.2866	84.747	48.5277	77.2648	78.4308	72.6383	25.2115
Glass	66.7814	73.9048	75.1515	15.0152	67.6255	69.9502	49.4459	35.513
Automobile	76.5643	81.4762	85.4571	29.3357	81.7667	74.5524	57.4143	32.7024
Breast Cancer	68.2734	73.1022	66.8879	71.8313	74.2808	72.8461	72.697	70.2956
Ionosphere	91.3127	91.6262	93.0476	92.1651	89.7444	87.0984	82.1675	64.1032
Horse Colic	80.6809	84.994	81.6299	84.6359	85.1554	79.1066	78.6997	63.0481

Table 3: True Positive, False Positive, True Negative and False Negative

Dataset	Classifier	TP	FP	TN	FN
Audiology	Neural Network	0	0.002233	0.997767	0.1
	Bagging	0	0	1	0.1
	AdaBoost	0	0.001324	0.998676	0.1
	Stacking	0	0.005316	0.994684	0.1
	J48	0	0	1	0.1
	IBK	0	0.000434	0.999565	0.1
	NaiveBayes	0	0	1	0.1
	ZeroR	0	0	1	0.1
Glass	Neural Network	0.777143	0.215333	0.784667	0.222857
	Bagging	0.774286	0.13381	0.86619	0.225714
	AdaBoost	0.79	0.13419	0.86581	0.21
	Stacking	0.117143	0.079524	0.920476	0.882857
	J48	0.714286	0.159048	0.840952	0.285714
	IBK	0.752857	0.156048	0.843952	0.247143
	NaiveBayes	0.744286	0.405238	0.594762	0.255714
	ZeroR	0	0	1	1
Automobile	Neural Network	0	0	1	0
	Bagging	0	0	1	0
	AdaBoost	0	0	1	0
	Stacking	0	0.2	0.98	0
	J48	0	0	1	0
	IBK	0	0	1	0
	NaiveBayes	0	0	1	0
	ZeroR	0	0	1	0
Breast Cancer	Neural Network	0.7926	0.5758	0.4242	0.2074
	Bagging	0.929	0.7368	0.2632	0.071
	AdaBoost	0.789	0.615	0.385	0.211
	Stacking	0.8938	0.6961	0.3039	0.1062
	J48	0.9473	0.74	0.26	0.0527
	IBK	0.8935	0.6613	0.3388	0.1065
	NaiveBayes	0.851	0.5663	0.4338	0.149
	ZeroR	1	1	0	0
Ionosphere	Neural Network	0.7958	0.0213	0.9787	0.2042
	Bagging	0.8103	0.0244	0.9756	0.1897
	AdaBoost	0.8544	0.027	0.973	0.1456
	Stacking	0.8406	0.0328	0.9672	0.1594
	J48	0.8207	0.0596	0.9404	0.1793
	IBK	0.6888	0.0272	0.9728	0.3112
	NaiveBayes	0.8646	0.2025	0.7975	0.1354
	ZeroR	0	0	1	1
Horse Colic	Neural Network	0.8517	0.2701	0.7299	0.1483
	Bagging	0.9259	0.2797	0.7203	0.0741
	AdaBoost	0.8649	0.2667	0.7333	0.1351
	Stacking	0.9189	0.2777	0.7223	0.0811
	J48	0.9307	0.2835	0.7165	0.0693
	IBK	0.8321	0.2791	0.7209	0.1679
	NaiveBayes	0.8013	0.2375	0.7625	0.1987
	ZeroR	1	1	0	0

An analysis of the accuracy results of the various datasets gives us the following information.

- **Multi-Class Datasets** - For the Audiology data set all 7 algorithms selected performed better than the base algorithm (ZeroR). The algorithm AdaBoost was found to be the most accurate performer followed closely by NN. For the Glass Identification dataset all the algorithms except for Stacking performed better than the base algorithm (ZeroR). The algorithm AdaBoost was again found to be the most accurate performer followed by Bagging. For the Automobile data set again all the algorithms except for Stacking performed better than ZeroR our base algorithm. Once again AdaBoost was found to be the most accurate performer followed by J48 and Bagging respectively.

- **Two-Class Datasets** - For the Breast Cancer dataset, the NN and AdaBoost algorithms performed worse than our base algorithm ZeroR. The most accurate performer was found to be J48 followed by Bagging and IBk respectively. For the Ionosphere dataset all the algorithms performed better than our base algorithm ZeroR. AdaBoost was found to be the most accurate algorithm, followed closely by Stacking and Bagging respectively. For the Horse Colic dataset again all the algorithms performed better than our base algorithm ZeroR. For this dataset J48 was found to be the most accurate algorithm followed closely by Bagging and Stacking respectively.

The confusion matrixes were built using the True Positive Rate, False Positive Rate, True Negative Rate and False Negative Rate. This shows us how accurate the model's predictions were, and from the results we can see exactly where things may have gone wrong. Apart from getting to know the counts of the actual versus predicted class rates, the confusion-matrixes are also useful in understanding the rate of Type I and Type II errors for the various algorithms and datasets.

In summary, we can tentatively conclude from the tabulated accuracy and confusion-matrix results that for a multi-class nominal dataset, AdaBoost should be the most preferred algorithm and Stacking should be the least preferred algorithm. From our analysis we can state that for the multi-class numeric dataset, AdaBoost should be the most preferred algorithm and Staking should be the least preferred algorithm. Similarly, for the multi-class mixed dataset, AdaBoost should be the most preferred algorithm and Stacking should be the least preferred algorithm. For all multi-class datasets the AdaBoost algorithm is consistently the best performing algorithm and stacking is consistently the worst performing algorithm.

From our analysis of the two-class nominal dataset, we can see that J48 should be the most preferred algorithm and AdaBoost should be the least preferred algorithm. For the two-class numeric dataset AdaBoost should be the preferred algorithm and Naïve Bayes should be the least preferred algorithm. The analysis for the two-class mixed dataset shows that J48 is the best performing algorithm and Naïve Bayes should be the least preferred algorithm. In the case of the two-class dataset no one algorithm is consistently the best or worst performer, but from the results we see that J48 performs well for the majority of the tested two-class datasets and that the Naïve Bayes algorithm performs poorly for most two-class datasets.

DISCUSSIONS AND CONCLUSIONS

With advances in information technology, the use and benefits of data mining has grown in multiple folds. For example, online companies such as Netflix can mine its customer databases to recommend rentals to individual customers based on their rental history. Financial institutions can suggest products (e.g., credit cards) to their customers based on analysis of their monthly purchasing power. Even small companies, such as local video stores or restaurants can mine their customer database to customize their offerings. However, companies could encounter challenges when utilizing data mining, especially in selecting an appropriate model and algorithm, due to a lack of knowledge and/or resources. The recommender systems that we proposed in our study could benefit managers/companies encountering this problem. For instance, let us consider that a company has a new customer dataset that is a two-class nominal type dataset, but has little knowledge regarding the appropriate model and algorithm for this dataset. Companies can overcome this problem by testing all the models and algorithms and selecting the best performing one. However, this approach could be very costly and time consuming, and might result in significant losses due to the delays involved in accurately analyzing the dataset. By using a recommender system, companies can resolve this issue. The data analyst can input the attributes of the dataset into the recommender system which would then search the result repository (F) and find that J48 was the best performing algorithm for this type of dataset. The recommender system would then suggest that the company use a decision tree model, specifically algorithm J48 for accurate analysis of the new customer dataset.

From a theoretical point of view, we tackled the problem of identifying the best algorithm given the dataset characteristics and utilizing a recommender system to perform the selection process. Our study extends prior studies by incorporating multiple performance measures and also by proposing a methodology of using the comparison results to select the most appropriate algorithm for a new problem. Furthermore, the results of our experiments indicate that AdaBoost is a relatively stable performer compared to other algorithms. The results of this study also indicate that the characteristics of datasets have an influence on algorithm performance. From a managerial point of view, this study shows that the performance of an algorithm varies with the dataset, and understanding this is critical for selecting the correct model

and algorithm. It also highlights the importance of the model and algorithm selection process and illustrates how a recommender system can not only lead to both cost and time savings, but also improve accuracyby potentially reducing errors in the model and algorithm selection process.

A limitation of our study is related to the dataset selection, i.e. the types of datasets that we selected were limited. This study has to be replicated with a larger sample of datasets to come up with a more generalizable result. Another limitation of this study is that the choice of models and algorithms was limited by Weka. Further work needs to be done with other models and algorithms in order to generalize the findings of this st... Also the results were compared only on the basis of accur... and error rate; inclusion of other comparative measures, s... as ROC, training time, testing time etc. can lead to ad... reliability of the findings. Finally, the use of a recommen... system is presented only at a conceptual level; future rese... work should try to develop a prototype to test the use o... recommender system for data mining purposes. Overall, ... paper illustrates the importance of selecting an appropr... model and algorithm and the need for having a recommen... system to assist managers in choosing the right model a... algorithm for any given classification problem.

REFERENCES

1 Adriaans, P. and Zantinge, D. (1996) Data Mining, Reading Mass.: Addison-Wesley.

2 Aha, D.W. (1992) Generalizing from Case Studies: A Case Study. In 9th Int. Machine Learning Conference

3 Bigus, J.P. (1996) Data Mining with Neural Networks, New York: McGraw-Hill, pp. 131-177

4 Brodley, C.E. (1995) Recursive Automatic Bias Selection for Classifier Construction. Machine Learning, 20:63-94

5 Brodley, C.E. and Smyth, P. (1997) Applying Classification Algorithms in Practice. Statistics and Computing

6 Chen, M.S., Han, J. and Yu, P.S. (1996) "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, pp.866-...

7 Donato, J.M., Schryver, J.C., Kinkel, G.C., Schmoyer Jr., R. L., Leuze, M.R. and Grandy, N. W. (1999) Mining Multi-Dimensional Data for Decision Support", Fut... Generation Computer Systems, Vol. 15, No. 3, pp. 433-441

8 Fayyad, U. M, Piatetsky-Shapiro, G. and Smyth, P. (1996) From Data Mining To Knowledge Discovery In Databases, AI Magazine, Vol. 17, pp.37-54.

9 Fayyad, U. and Stolorz, P. (1997) Data Mining and KDD: Promise And Challenge, Future Generation Computer Systems, Vol. 13, No. 2-3, pp.99-115

10 Fayyad, U., Piatetsky-Shapiro, G., Smith, G. & Uthurusamy, R. (1998). Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press.

11 Gama, J. and Brazdil, P. (1995) Characterization of Classification Algorithms. In Pinto Fereira, C. and Mamede, N. Editors, Progress in AO, 7th Portuguese Conf. ... (EPIA'(%), pages 83-102. Springer Verlag

12 Gordon, F. and DesJardin, M. (1995) Evaluation and Selection of Biases. Machine Learning, 20:5-22 (http://www.ai.univie.ac.at/oefai/ml/metal/metal-theoretical.ht...

13 Hui, S.C. and Jha, G. (2000) "Data Mining For Customer Service Support", Information and Management, Vol.38, No. 1, pp. 1-13

14 Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In C.S. Mellish (ed.), Proceedings of the Fourt... International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, pp. 1137-1143.

15 Kohavi, R. (1996) Data Mining using MLC++, A Machine Learning Library in C++. In Tools with AI

16 Linhard, H. and Zucchini, W. (1986) Model Selection. NY: Wiley

17 Mitchel, T. (1997) Machine Learning. MacGraw Hill

18 Nakhaeizadeh, G. and Schnable, A. (1997) Development of Multi-Criteria Metrics for the Evaluation of Data Mining Algorithms. In KDD'97, pages 37-42. AAAI Press.

19 Pitta, D. (1998) "Marketing on-to-one and Its Dependence on Knowledge Discovery in Databases", Journal of Consumer Marketing, Vol. 15, No.5, pp. 468-480

20 Provost, F.J. and Buchanan, B.G. (1995) Inductive policy: The Pragmatics of Bias Selection. Machine Learning, 20:35-61, 1995.

21 Rendell, L.A., Seshu, R.M., and Tcheng, D.K. (1987) Layered Concept Learning and Dynamically Variable Bias Management. In 10th Int. Joint Conf. on Art... Intelligence, pages 308-314

22 Shavlik, J.W., Mooney, R. and Towell, G. (1991) Symbolic and Neural Computation: An Experimental Approach. Machine learning, 6:111-114

23 Salzberg, S. A. (1991) Nearest Hyper-Rectangle Learning Method. Machine Learning, 6:251-276

24 Schaffer, C. (1993) Selecting a Classification Method By Cross-Validation. Machine Learning, 13:135-143

25 Schaffer, C. (1993) Selecting a Classification Method by Cross-Validation. Preliminary Papers of the Fourth International Workshop on Artificial Intelligence and Stat... (pp. 15-25)

26 Sinha, A.P., and May, J.H. (2005) Evaluating and Tuning Predictive Data Mining Models Using Receiver Operating Characteristic Curves. Journal of Manage... Information Systems, 21, 3, 249-280.

27 Solomon, S., Nguyen, H., Liebowitz, J., and Agresti, W. (2006). Using Data Mining to Improve Traffic Safety Programs, Industrial Management & Data Systems, Vol. 1... 5, 621-643.

28 Sung, H.H. and Sang, C.P. (1998) "Application of Data Mining Tools to hotel Data Mart on the Intranet for Database Marketing", Expert Systems With Applications, Vo... No. 1, pp. 1-31

29 Weiss, S.M. and Kapouleas, I. (1989). An Empirical Comparison of Pattern Recognition. Neural Nets and Machine Learning Classification Methods. In 11th Int. Joint C... in Artificial Intelligence

30 Witten, I.H., and Frank, E. (2000) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation. San Francisco, CA: Morgan Kaufmann.